Issues for a Linux 'L4' trigger farm

Heidi Schellman

April 25, 2001

Introduction

This note examines some of the issues related to a 'L4' backend trigger farm at D0. The basic idea is:

- The existing L3 NT farm is used for detector readout and event building.
- Instead of running L3 filters on those nodes, the data are copied over ethernet connections to a farm of 'L4' nodes which run the existing L3 filter algorithms.

I assume that, in the long run, there are 48 DAQ nodes and that the maximum initial input event rate is 1000 Hz. In this case, each machine receives ∼40 250 KB events/second or an average rate of 5 MB/sec. In the 'L4' model, this full data rate must be transferred to the 'L4' system.

The L4 system is assumed to consist of at least 48 nodes.

1 Implementation Issues

There are 3 major issues in implementing such a system.

1. The feasibility of reading in and writing out 5 MB/sec from the DAQ nodes must be demonstrated. We know from the enstore mover nodes that transfer rates of over 10 MB/sec are possible for commodity PC's receiving data from tape drives and transferring it to other PC's. However both the hardware and operating system are different in this case.

We know from the enstore tests that the receiver L4 PC's can handle these rates, and, if the L4 farm expands, the rate per PC actually decreases, making the problem easier.

This can be demonstrated in bench tests.

In the long run, we may need to use GB ethernet as well as 100MB on the DAQ nodes as input event rates increase to 2000 Hz or above.

2. The purchasing/integration of the L4 farm itself. We have substantial experience in doing this for offline farms. Current experience indicates that purchasing/installation requires $\sim$ 1 FTE for 1-2 months and can be shared with regular farm purchases. Costs per dual node, including networking costs should be around 3K$. These nodes come rack-mounted, 20/rack and each rack consumes (and dissipates) a maximum of 6kW. System administration tools allow nodes to be 'cloned' over the network, which makes upgrades/patches very easy.

   A 48 processor farm would cost 150K$, a 100 processor farm would cost 300 K$. Offline experience indicates that maintenance requires $<$ 0.5 FTE.

   For tests, 20 dual 850 MHz nodes from the existing offline farms could be installed at D0 on a short time scale.

3. The control software for the data transfers and L4 systems. Much of this has already been written for the existing L3/DAQ system but would have to be ported to Linux. There is also considerable expertise building such systems from other Hep experiments which may be usable.

   We already have some experience with configuration control on distributed linux farms. In addition, the Fermilab farms group and the d0 systems group have several general monitoring tools, notably displays of resource utilization on farms.

2 Performance

Building a 'L4' system provides two things:

1. The physics algorithms run under a more familiar operating system and configuration control is easier.

2. The event building and filtering are decoupled. This makes it possible to optimize them separately. For example, if new faster CPUs become available, they can be added to the L4 farm without modification to the specialized DAQ nodes. Conversely, the DAQ nodes can be improved independently of the L4. This makes the whole system much more scalable as Run II progresses.

I looked at the CPU/event costs of 2 L4 farms:

- a 48 node farm would cost $\sim$ 150K$. The CPU available per event, assuming that reading data in on the NT nodes takes 0.3 of a CPU at 10 MB/sec and network transfers take 0.5 of a CPU at 10 MB/sec is around 85 msec. This is not very different from the 90 msec CPU/event available in the present system.

- a 48 node upgrade to that system would cost $\sim$ 150K$. The CPU available per event grows by more than a factor of 2 as all the additional CPU is available for algorithms.

The bottom line is that a 50 node L4 farm buys you convenience and future scalability but not more CPU, but after that, one can replace or increase the CPU of the L4 farm just by adding more nodes.

3 Possible Plan

To study this further we would need to:

- Start bench tests with an existing L3/DAQ node (or an NT box which can emulate a fully loaded L3/DAQ node) and a linux box to understand the data transfer issues at top rate.

- Come up with a strawman proposal for the software needed for configuration control and event accounting in a L3/L4 system. Estimate the effort required and the resources available.

- Get a small (20 processor) L4 system into DAB for real tests.